DOCUMENT RESUME

ED 267 127 TM 860 220

AUTHOR Dovell, Patricia; Buhr, Dianne
TITLE Essay Topic Difficulty in Relation to Scoring Models.
PUB DATE Apr 86
NOTE 43p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS College Students; *Difficulty Level; Essays; *Essay Tests; Generalizability Theory; Higher Education; Interrater Reliability; Latent Trait Theory; *Models; Regression (Statistics); Research Design; *Scoring; Statistical Analysis; *Test Theory; Writing Evaluation

ABSTRACT

This study examined the difficulty level of essay topics used in the large-scale assessment of writing in relation to five different scoring models, and sought to determine what effects the scoring models would have on passing rates. In model one, examinee's score is the direct result of a score assigned by the reader or the sum of scores assigned by multiple readers. Model two reports a composite writing score which combines a score from a direct assessment with a score from an indirect assessment. Model three adjusts the essay score or combined score by using the objective measure to compute the regression coefficient. Model four scores polychotomous items, such as essays (e.g. Rasch Model). Model five corrects raw scores for differences among raters adjusted by generalizability theory. Models one and two produced essentially equivalent pass/fail decisions. Model three was found to be inappropriate for making decisions about individual students because of the essay scores' discrete scale. Models four and five could not be applied because our data did not fit the design. Results showed that each of the models has its advantages depending upon the purpose of the assessment and the nature of the data. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED267127

ESSAY TOPIC DIFFICULTY IN RELATION TO SCORING MODELS

by

Patricia Dovell

Dianne Buhr

University of Florida

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. Dovell

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Paper presented at the annual meeting of the
American Educational Research Association
San Francisco, April, 1986

TM 860 220

A B S T R A C T

This study examined the difficulty level of essay topics used in the large-scale assessment of writing in relation to five different scoring models, and sought to determine what effects the scoring models would have on passing rates. Models 1 and 2 (which were models of direct assessment and combined score) produced essentially equivalent pass/fail decisions. Model 3 (a regression model) was found to be inappropriate for making decisions about individual students because of the essay scores' discrete scale. Models 4 and 5 (which were models of polychotomous item response theory and rater reliability adjusted by generalizability theory) could not be applied because our data did not fit the design. Results showed that each of the models has its advantages depending upon the purpose of the assessment and the nature of the data.

The difficulty level of essay topics used in the large-scale assessment of writing is an issue of concern to the testing and measurement profession. In order for valid and reliable decisions to be made about examinees over different administrations of an essay examination, scores must be comparable though topics may vary in difficulty.

Objectives

To study the difficulty level of essay topics, our objectives are to:

- define difficulty level, both abstractly and empirically,
- review the literature to acquire historical perspective, and to identify the scoring models in use today,
- examine the data from a large-scale writing assessment, and manipulate the data to see the effect of alternate scoring models on passing rates,
- discuss the conclusions which resulted from this manipulation, and
- propose areas of inquiry for future research.

Theoretical Framework

We reviewed the literature to determine the state of the art in assessing the difficulty level of essay topics and to identify models which have been developed to control for variance in difficulty between multiple topics for a single administration as well as equating topics over multiple administrations. In our review we identified the following questions:

- What is difficulty level? How is it defined?
- What are the sources of variance in item difficulty and in examinees' scores?
- What are the scoring models?
- How do the scoring models control for variance in difficulty?

In our search, we discovered that the literature on essay topic difficulty is sparse. As pointed out by Breland (1983, p. 19), "The history of direct writing skill assessment is dominated by the issue of reliability...."

Given that the literature on difficulty level is sparse, the issue nevertheless is discussed indirectly. Hoetker (1982) discusses the factors that affect the difficulty of the essay topic, and also explains why the importance of this subject is now emerging. National testing services developed essay topics to measure achievement. Their goal was to spread out the examinee scores across a continuum. However, the competency tests, including our data source, seek to distinguish between two groups of writers--those who are judged to be competent writers and those who are judged to lack competent writing skills. When dealing with competency tests, the essay topic becomes much more crucial because of the need for the test to make accurate distinctions.

Hoetker (1982, p.381) convincingly makes the point that essay topic difficulty is worthy of further research attention when he reports a situation that occurred in California:

"To cite just one example: students taking the California State University and Colleges Equivalency Examination write a 90-minute essay on a set topic. The topics for the 1974 examination were provided by a committee of English professors, with the details 'explicitly left to the discretion of the Committee on English.' When scores on the 1974 examinations were found to differ drastically from those on the 1973 examinations, the topics were reviewed. It was concluded that the 1974 topic, which called for highly abstract reasoning, was manifestly more difficult than the previous year's topic, which called for reflection upon a personal experience. The report of the affair concludes, 'It is the intention of the directors...to give more thought, more attention, and more money to the development of essay questions.'"

Generally the difficulty level of an essay topic is assumed to be represented by the mean score on responses to that topic. This definition has several drawbacks. First, the mean does not reflect just the difficulty of the topic. In addition to measurement error, the mean is a function of the sample of examinees, essay topics, and readers, plus the possible interaction of readers with topics and with examinees.

Secondly, given the testing situation in which the student chooses to write on one of two topics, the mean is affected by the students who chose to write on that topic. Rosenbaum (1985, p.14) identifies this problem in his discussion of the College Board's 1982 Advanced Placement Examination in Biology:

"The students selecting one essay rather than another may differ systematically, as was the case with the essay pair consisting of Essay #5 and Essay #6. In particular, the mean number of the multiple choice items answered correctly was 70.66 for the 4129 examinees selecting Essay #5 and 65.23 for the 11547 examinees selecting Essay #6, with a two-sample t-statistic of 16.7."

The idea is that a student should not benefit if he chooses to write on a topic selected by the least able examinees, and he should not be penalized if he selects the topic also selected by the most able examinees.

As an interesting aside, Meyer's (1939) research revealed that students could not select the essay topics on which they would score best. Meyers had students write essays in response to five questions, and asked them to identify one of their five responses to be omitted from scoring. He found that (p. 164) "...forty percent [of one group] and ...forty-six percent [of the second group] made better scores when the answer which would have been omitted was counted in place of the poorest of their four choices."

A third problem with defining difficulty level as the mean is that the numerical value is not directly interpretable. Given a scale of 2 through 8, is an essay topic with a mean of 5.1 more difficult than a topic with a mean of 5.3? To answer this question, the standard error of measurement of the mean needs to be calculated, and confidence intervals need to be identified to determine if the differences between the means are statistically significant.

Another problem that researchers face when looking at difficulty level defined as the mean is that students' scores generally cluster around the mean. Hoetker (1982, p. 378-379) explains the statistical limitations of these data:

> "This 'piling up' of holistic ratings is probably unavoidable, short of major changes in the conventions of the process of scoring (such as, for example, directing raters to assign a predetermined percentage of essays to each quality category). This piling up has two important consequences which researchers generally have not recognized. First, there is typically so little variance among ratings that any study using them is almost guaranteed to yield results of 'no difference.' How this works is illustrated in the discussion below. Second, the fact that holistic ratings pile up around the mean suggests that in most cases parametic statistics, such as t-tests and analysis of variance, are inappropriate, since such methods assume an underlying normal distribution of scores."

In summary, our review of the literature showed that essay topic difficulty level is not directly defined. It is generally assumed to equal the mean score of the item responses, but there are problems with this assumption, as demonstrated above.

We can define the difficulty level to be the condition which allows the examinee to demonstrate his writing ability. However, until this definition can be translated into operational measurement terms, it will not be fruitful.

Prior to identifying the models that have been developed to control for variance in topic difficulty, it is helpful to review the sources of variance in examinee scores. These sources of variance are due to differences in examinees, differences in raters, and differences in essay topic difficulty, plus error. Our task is to identify scoring models, and see how they control for variance in essay topic difficulty.

The classic model, which we will call Model 1, is the direct assessment. This model is politically and academically popular, probably because it is easily understood and appeals to our notion of common sense. Diederich (1974, p.1) explained the popularity of the direct assessement model with the analogy that "Whenever we want to find out whether young people can swim, we have them jump into a pool and swim."

In this model the examinee writes an essay or multiple essays, his essay is read (holistically or analytically) and assigned a score within a range (such as 1-4). The examinee's score is the direct result of the score assigned by the reader or the sum of the scores assigned by multiple readers. Meredith (1984, p.14) explains how this works:

"Basically, student scores are derived by one of two procedures: (1) summing or averaging the values assigned by two raters and resolving differences if the ratings are on different sides of the cut point or (2) summing or averaging the ratings of two or three raters and resolving any discrepancies greater than one point."

To increase the reliability of this scoring model, the literature strongly favors three conditions:

1. all examinees write on the same topic;
2. multiple writing samples be collected from each examinee; and
3. multiple readers read each writing sample.

In addition, other conditions are often recommended (i.e., that split ratings be adjudicated, that readers be trained, and that readings be conducted with readers assembled on-site in one location).

In spite of these recommendations, many large-scale assessments do not follow these conditions. Why not? First, it is politically popular to present examinees with a choice of topics. Secondly, for reasons of test security, new topics must be introduced for each administration. Third, collecting and scoring multiple writing samples is both time-consuming and costly. Multiple-readers are generally used, although the number of readings per writing sample varies across programs.

How does Model 1 control for the variance in difficulty level? Testing programs control for it by pilot testing prompts. Williams (1983, p.8) recommends this method when he writes, "One [way] might be to simply field test as many prompts as possible and then only use those operationally that produce similar score distributions." Freedman and Robinson (1982, p.396) reiterate this position when they outline criteria for essay topics, "...most important, topics must discriminate accurately between good and poor writers. Our only measure of how well the test discriminates comes from our pretesting." Using the results of the pretest, this model pre-equates the difficulty level of essay topics by rejecting topics that do not elicit suitable and comparable score distributions.

Model 2 reports a composite writing score which combines a score from a direct assessment (writing sample), with a score from an indirect assessment (multiple choice format). How this model works is described by the College Board in its explanation of the scoring for the General Examination in English Composition (1981, p.6):

"...Each essay is read and graded by two professors, the sum of the two grades is combined with your multiple-choice score, and the result is reported as a scaled score between 200 and 800.

CLEP does not report separate scores for the multiple-choice section and essay section of the general Examination in English Composition for two reasons. First, neither of the two sections alone is sufficient to assess reliably a candidate's writing skills. Second, although the format of the two sections is very different, both the multiple-choice format and the essay format measure essentially the same type of ability--that is, expository writing skills."

The composite score model is also used in the Advanced Placement Program (APP) Examination in English. In this program, "The candidates' composite score--multiple-choice plus essay--is reported as a grade of 1 (low) to 5 (high), with credit recommended for grades of 3, 4, and 5." (College Entrance Examination Board, 1977, p.2).

In order for the composite-score model to be valid, it must be demonstrated that both the direct assessment and the objective assessment measure the same skills. When this is the case, the composite score is often preferred over the direct assessment score because two types of measures combined have been shown to be more reliable than one (Godshalk, Swineford, & Coffman, 1966).

Tied to the validity concern is the task of equating essay topics. Breland (1983, p. 19-20) states the problem and the usual solution:

"A validity issue for which no evidence was found is that related to the equating of essay assessments. Since topics and specific tasks vary in difficulty, and since each administration of a test must necessarily change the topic for security purposes, a not inconsequential problem is how best to equate a score received in one administration with a score received in another. This

problem is usually handled through a combined essay and objective assessment in which the equating is performed on the combined score using an objective measure. However, if an essay assessment were used in isolation, it is not immediately apparent how equating across administrations could be achieved."

This discussion seems to suggest that the variance in essay topic difficulty is diminished through the composite-score scoring model, because the essay score is only one component of the score, and the procedures for equating objective measures are well-established.

Model 3 is the regression model in which the reported score is adjusted through the use of regression. In order for the use of regression to be valid, it is necessary to have a highly-correlated covariate. In the case of an essay examination administered in conjunction with an objective writing test, the objective measure serves as the covariate.

The regression model can take one of two forms: the essay score or a combined score can be adjusted using the objective measure to compute the regression coefficient.

The procedure underlying the regression model is explained by Hills (1972, p. 139) in the following example:

"Consider giving Form A of a final examination to the Fall Quarter class, and Form B to the Winter Quarter class. If one has included in each form a set of items, call them Part X, so that a score on those items can be obtained for each student in both classes, one makes the assumption that the regression of final examination score on X score is the same in both groups. He then computes the regression coefficient in the group to which he wishes to equate."

Although Hills does not address the scoring of essays directly, he does give multiple examples of common material that may be used as an "anchor variable to equate the mean and standard deviation" (p. 140). Hills' examples include "a common final examination, common admissions or academic-potential variables, or anchor sections of common material in examinations" (p. 145).

The regression model, unlike models 1 and 2, recognizes variance in essay topic difficulty, and equates scores through the mathematical manipulation of a second measure. In contrast, models 1 and 2 handle variance in essay topic difficulty "on the front-end" by pilot testing, and reducing the pool of acceptable essay topics to those that produce comparable and desirable scoring distributions.

Model 4 is the item response theory model. As Williams (1983, p. 2) points out, "In many proficiency programs, Item Response Theory (IRT) methods have been adopted for the equating designs, in particular the one-parameter (1PL) model." The 1PL model, also known as the Rasch model, is a well-known scoring model frequently applied to dichotomous items. The Rasch model, in fact, is used to score the objective subtests of our data source. Recently the model has been extended to score polychotomous items, such as essays.

When this scoring model is applied to polychotomous data, results such as those cited by Williams (1983, p. 5) may emerge:

"Score Distributions of Three Writing
Prompts Scored Holistically
and 1PL Difficulty Estimates

Score Point	Prompt 1 (.12)	Prompt 2 (-2.86)	Prompt 3 (.49)
4	15%	56%	10%
3.5	7	18	10
3	29	15	22
2.5	7	2	14
2	31	5	22
1.5	1	.3	6
1	6	.9	14
NSR	4	3	.9"

Notice in Williams' example that Prompt 2 elicited responses in which 56 per cent of the essays received the maximum score, whereas Prompt 3 evoked only 10 per cent of the top scores. Clearly prompts which produce such dramatic differences in ratings need a scoring model that equates for essay topic difficulty.

De Gruijter (1980, p. 248-249) also cites item response theory (also known as latent-trait theory) as a method to equate for differences between essay topics:

> "The method of free choice of questions seems to be more appropriate when one is interested only in writing abilities. When subject-matters are tested a free choice would lead to the undesirable possibility that students can make a good impression because of this choice. Under these circumstances only a choice among difficulty levels of questions, not among questions themselves, seems allowed; in British examinations choice among questions is necessary because curricula are allowed to differ. This procedure needs a system to adjust for the differences between questions. This is done by equating. When scores on different questions are non-linearly related, equating is most conveniently realized within the framework of latent-trait theory."

It is interesting to note that de Gruijter sees the greatest usefulness in the IRT model to be realized when the focus of the test is upon the content of the essay response, rather than writing ability.

Model 5 is a model in which raw scores are corrected for differences among raters. In order for this model to be valid, it is necessary to identify the variance due to raters. This requires the use of generalizability theory. Before focusing directly on model 5, it is useful to review the main idea underlying generalizability theory.

Generalizability theory examines the measurement conditions, called facets (which in our case are raters and topics), and uses analysis of variance to separate variance due to different facets. The rationale behind this approach is explained by Shavelson (1984, p. 115-116):

> "Reliability, then, refers to the generalizability or dependability of scores (Cronbach et al., 1972). As the number of facets entering into a measurement...increases, the possibility of introducing error into the measurement increases. As error increases, the generalization from a single score to the...average score may become increasingly less reliable. That is, as the number of facets of a measurement increases, the number of potential sources of error in a measurement increases. Increasing error creates increasingly unreliable measurements.
>
> A measurement model, then, should estimate the magnitude of error introduced into a measure of unit performance by each facet. Moreover, it should provide information on how to reduce that error in the most cost-effective way. Generalizability theory provides the basis for accomplishing this."

To examine the sources of variance, generalizability theory can be used to conduct a generalizability (G) study, and a decision (D) study. Ideally a generalizability study should be conducted first to provide information about the question of interest before the decision study is used to make decisions (such as pass/fail decisions) about members of the population.

Our question of interest is what effect do the topic, the rater, and the interaction of the two have upon the reliability of the examinee's score. Ideally this question could be addressed with a generalizability (G) study. Our measurement conditions would have two facets--topics and raters. The facets would be random, since they refer to a "more extensive set of conditions for the universe of generalizability" (Crocker & Algina, p. 188), namely all topics that meet the specifications and all raters who fulfill the requirements. Although a number of designs are possible, two designs are suggested below.

Design 1 would be a fully crossed design. In this design the two facets are "crossed", that is,"...every measurement condition of the first facet occurs in combination with every measurement condition of the second factor" (Crocker & Algina, p. 178). In our example each examinee (P) would write on each topic (T) and have each essay scored by each rater (R).

Design 1 is visually represented in Figure 1.

Figure 1

	T 1					T 2				
	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5
P1										
P2										
P3										
.										
.										
.										
.										
.										
.										
.										
P100										

This figure represents the situation in which 100 persons (P) write on two topics (T) each. Every essay will then be rated by five raters (R).

A second possible design, which we'll call Design 2, is a partially nested design. In this design each examinee writes on each topic. Then the essays for a set of examinees are rated by a set of raters, while the essays for the remaining examinees are rated by a second, mutually exclusive set of raters.

Design 2 is visually represented in Figure 2

Figure 2

	T 1			T 2		
P1	R1	R2	R3	R1	R2	R3
P2	R4	R5	R6	R4	R5	R6
.						
.						
.						
.						
.						
.						
.						
P100						

This figure represents the situation in which 100 persons (P) write on two topics (T) each. Each examinee's two essays will then be rated by a set of three raters (R). The raters are members of mutually exclusive sets.

Both the fully crossed design and the partially nested design have the premise that each examinee writes on both topics.

De Gruijter (1980, p. 250) comments on the need to understand the design of a study, and to use this scoring model only when it is justified:

"....it is not clear whether students' raw scores or scores corrected for mean differences among raters should be used in the assignment of marks. If the scores are not corrected, raw scores will differ partially because of rater differences in leniency. On the other hand, using raw score deviations from the various means probably leads to overcorrecting for rater variation. One could avoid this uneasy situation by means of another design for rating the answers. For instance, one could assign different questions to different raters."

De Gruijter's article continues to reference several methods that correct for rater bias. The methods require the use of analysis of variance or covariance, and are only appropriate for particular designs. His article also cites an important distinction made between two kinds of rater bias (p. 259):

"Krippendorff (1970) defines two kinds of bias in ratings: the bias can consist of a tendency to use only one side of the scale (resulting in a deviation of the mean score from the grand mean) or of a tendency to use a larger or smaller part of the scale (affecting the score variance)."

Here again the appropriate correction method relates to the nature of the design.

De Gruijter's highly technical article includes an important observation (p. 250-251):

"In generalizability theory such problems [rater bias] can easily be recognized in advance. Sometimes a preliminary study can be done beforehand in order to obtain an estimate of all relevant variance components."

What de Gruijter seems to be recommending is preventive medicine. In other words, it is better to control for rater bias than to adjust for it with a scoring model. Although a discussion of rater reliability is beyond the scope of this paper, there is abundant literature on the topic, which emphasizes scoring criteria and rater training.

Scoring models 3, 4, and 5 each incorporate equating methods with the goal of increasing the reliability of scores. Meredith and Williams (1984, p. 15), in their thoughtful discussion of issues facing the direct assessment of

writing, point to a problem inherent in the narrow range of scores assigned during holistic scoring:

"The test score equating analog for direct writing assessment is not quite as clear as it is in indirect assessment. The applicability of equating direct assessments is compromised by the discreteness of the score scale. In programs where two raters evaluate student responses on one prompt, the score scale may only range from one to four at half-point intervals. A scale of this type may not provide a sufficient number of score points to use in the equating process."

The results of equating for essay topic difficulty when dealing with a narrow range of raw score data will be examined later in this paper when various scoring models are applied to our data source.

Having presented five unique scoring models, we will now proceed to a description of the methods.

Methods

Five scoring models are identified. Model 1 sums the raw scores assigned by readers to create a scaled score. Model 1 is the actual scoring model used in the large-scale assessment. Therefore the particulars of this method are described in the "DATA SOURCE" section of this paper.

Model 2 reports a composite writing score which combines a score from a direct assessment (writing sample), with a score from an indirect assessment (multiple-choice format). Combined scores for our data were calculated by the following method:

The score from the direct assessment was converted to the same scale as the indirect assessment by using the linear conversion of 30X + 300. The values of 30 and 300 were used because 300 is the approximate mean and 30 is the approximate standard deviation of the indirect assessment. A 30 point

difference between scores was selected because this requires an examinee who fails one subtest to score at least one standard deviation above the mean on the other subtest in order to pass by the combined score.

The scale was shifted so that a raw score of 5 on the direct assessment would be equal to a scale score of 300. This was done so that the mean of the direct assessment (approximately 5) would be converted to the mean of the indirect assessment (approximately 300). The linear conversion of the shifted scale resulted in the following:

Essay Score	Converted Essay Score
2	210
3	240
4	270
5	300
6	330
7	360
8	390

The passing score for the composite scoring model was determined by adding the passing score required on the indirect assessment to the converted passing score on the direct assessment. A passing score of 4 is required on the essay, which is converted to 270. A passing score of 265 is required on the indirect assessment. Thus the passing score required for our composite scoring model is 535.

Model 3 adjusts scores to account for variability in essay topic difficulty. This model equates scores with the use of regression on another measure, in our case multiple-choice writing items.

Essay scores within and between administrations for three test administrations were equated by means of a procedure adopted by Hills (1972) from a method suggested by Gulliksen (1950). This procedure uses a common

examination to place scores from different essay topics on the same scale. This is done through the regression of essay scores for one topic on the score of the common examination. In our case the common examination was the multiple-choice writing subtest. This procedure assumes that the regression of the essay score on the multiple-choice writing score is the same for two groups taking different topics.

Topic 1 of the March 1983 assessment was arbitrarily selected as the base value, and the regression coefficient of Topic 1 scores on multiple-choice writing scores was obtained. This coefficient (.022398) was used to adjust the mean and variance for Topic 2 scores, according to the following procedure:

$$\bar{X}_{E2A} = \bar{X}_{E1} + b_1\left(\bar{X}_{W2} - \bar{X}_{W1}\right)$$

where $\bar{X}_{E2A}$ is the adjusted mean for Topic 2 expressed in terms of Topic 1

$\bar{X}_{E1}$ is the mean for Topic 1

b_1 is the regression coefficient for Topic 1 scores on the multiple-choice writing scores

$\bar{X}_{W1}$ is the mean on the multiple-choice writing subtest for examinees choosing Topic 1

$\bar{X}_{W2}$ is the mean on the multiple-choice writing subtest for examinees choosing Topic 2

$$\sigma^2_{E2A} = \sigma^2_{E1} + b_1^2 \left(\sigma^2_{W2} - \sigma^2_{W1} \right)$$

where σ^2_{E2A} is the adjusted variance for Topic 2 scores expressed in terms of Topic 1

σ^2_{E1} is the variance for scores on Topic 1

b_1^2 is the squared regression coefficient of Topic 1 scores on the multiple-choice writing subtest scores

σ^2_{W2} is the variance of the multiple-choice writing subtest scores for Topic 2 examinees

σ^2_{W1} is the variance of the multiple-choice writing subtest scores for Topic 1 examinees

The same procedure was followed in adjusting means and variances of essay topic scores obtained from two other 1983 test administrations. The adjusted means and variances were then used to obtain rescaled scores for individuals as follows:

$$X_{E2A} = \bar{X}_{E2A} + \frac{\sigma_{E2A}}{\sigma_{E2}} \left(X_{E2i} - \bar{X}_{E2} \right)$$

where X_{E2A} is the rescaled score for examinee i

$\bar{X}_{E2A}$ is the adjusted mean for Topic 2 scores

σ_{E2A} is the adjusted standard deviation for Topic 2 scores

σ_{E2} is the standard deviation of Topic 2 scores

X_{E2i} is the score for examinee i on Topic 2

$\bar{X}_{E2}$ is the mean score for Topic 2

The rescaled scores were then used to determine the percentage of examinees passing at the cutting score of 4 for each topic for the three test administrations.

In addition, the Pearson product-moment correlation between the essay scores and multiple-choice writing scores was calculated. This step was performed because a highly-correlated covariate is a necessary condition of the regression scoring model.

Model 4 is an item response theory model applied to polychotomous scores. Difficulty estimates for different topics are obtained from a field-test in which the same examinee writes on multiple topics. In our data we have only one writing sample per examinee. This did not permit us to establish difficulty levels for different topics. Therefore, because our data did not fit the design, we could not apply this scoring model to our data.

Model 5 is a rater reliability model in which scores are adjusted through the use of generalizability theory. Appropriate designs, such as the fully crossed design or the partially nested design, require an examinee to write on two or more topics. Since our data did not fit the design, we could not apply this scoring model.

Data Source

Data from a large-scale writing assessment were studied. The examination is administered three times a year (March, June, and September or October) to college students at the end of their sophomore year. Each student has 50 minutes to write an essay on one of two topics. The scoring criteria are provided to the student. Topics are changed with each administration. The scoring method is a modified holistic scoring, in which

scores are assigned with reference to range-finders. The readers are assembled on-site in one, two, or three locations. Readers are trained before and during the reading. Scores of one (low) through four (high) are assigned. Each essay is read twice. The reported score is the sum of the two ratings (2-8). However, split ratings (i.e. 1-3, 2-4, and 1-4) are refereed, and one of the two scores is replaced. Beginning with the Fall 1984 administration of the test, a passing score of 4 was established. Starting with the same administration, total scores of 3 were also refereed. This policy was established to correct the situation in which an essay was "failed" by one rater (score of 1) and "passed" by another (score of 2).

In addition, a writing subtest, objectively scored, is included in the assessment. Scores range from approximately 200 to 400, with a mean of approximately 300 and standard deviation of approximately 30. A score of 265 is required to pass.

The tables in the Appendix present descriptive data for six administrations. Tables 1 through 6 describe the essay statistics for each of six administrations. Scores on each topic are presented with and without referee correction. The mean, standard deviation, alpha, and number of examinees is listed, along with breakdown by race and sex. The number and percentage of split ratings is reported, along with the percentage of essays receiving each score.

Table 7 presents the difference in means for essay topics. The differences reported are a result of subtracting the topic two mean from the topic one mean. Positive values indicate that topic one had a higher mean. Negative values indicate that topic two had a higher mean.

Table 8 presents the difference in coefficient alpha for the essay

topics. The differences reported are a result of subtracting the topic two alpha from the topic one alpha. Positive values indicate that topic one had a higher alpha. Negative values indicate that topic two had a higher alpha.

Table 9 presents the percentage of examinees choosing each topic. It is interesting to note that, except for the March 1984 administration, there was a decided preference for one topic over the other.

Results

The results of applying the direct assessment scoring model to our data are reported in Tables 1 through 9. Within administrations, the largest difference between means for the two topics was .238 for October 1983. The average absolute difference in means for the six administrations was .147. Across administrations, differences were larger, with the greatest difference .508, between October 1983 topic one and March 1985 topic two. This is about one third of a standard deviation.

The results of applying the composite scoring model to our data appear in Tables 10 through 16. Table 10 describes the mean and standard deviation for the essay subtest, and the multiple-choice writing subtest. Table 11 presents the results achieved (mean, standard deviation, and N) when the composite score is calculated according to the method previously described. Table 12 contrasts the percentage of examinees passing using scoring model 1 with scoring model 2. The number and percentage of examinees passing the essay subtest and multiple-choice writing subtest are presented, along with the number and percentage passing both of these. For all six administrations there is a slight increase in the percentage who pass when the composite scoring model is used compared with the percentage who pass both subtests. Increases range from .6% to 4.4% of the examinee population. This means that

examinees who were near the cutting score but failed on one subtest may be pulled up to pass on the combined score by their performance on the other subtest.

Table 13 shows the number of examinees who passed by the combined score but failed either the essay or the writing subtest. For the first three administrations, about three times as many examinees failed the essay examination as those who failed the writing subtest. However, the numbers were about equal for the later administrations.

Table 14 shows the Pearson product-moment correlation between the essay scores and the objective writing scores. The correlations for the six administrations are consistent, ranging from .49 to .51.

Table 15 describes the essay means and standard deviations by topic, and also reports the multiple-choice writing subtest means and standard deviations for the groups that chose each topic. Table 16 presents the passing rate by topic for each of the two subtests, and for the composite score. In only one case does the composite score reduce the difference in passing rates between topics (March 1983). Where the difference is greatest between passing rates for essay topics (6.1% in June 1984), the difference for the combined score is even larger (7.0%).

The results of applying the regression scoring model to our data appear in Tables 17 through 19. Table 17 presents by topic the essay subtest's means and standard deviations adjusted by the regression equation presented in the "Methods" section of this paper. Adjusted means differ by at most .18. Table 18 presents by topic the rescaled essay scores for examinees. Table 19 compares the percent of examinees passing with and without the regression equation rescaling. There is a dramatic decrease in the percent passing for

the Topic 2, March 1983 group (17% decrease) and for the Topic 1, June 1983 group (22% decrease). This is because the regression equation shifted scores of 4 (the passing score) to scores of 3.681 and 3.965 respectively for those two groups. Conversely, no one who failed on the direct assessment scoring model passed with rescaled scores. This is because the increase in rescaled scores is insufficient to move examinees from one level to the next higher level.

Discussion and Conclusions

Model 1 is politically and academically popular. Many educators believe that testing writing ability with a writing sample will encourage the practice of writing in the curriculum. In addition, the ability of multiple-choice tests to measure writing ability is still regarded with suspicion. In fact, many multiple-choice writing items do focus on skills at the sentence level, and rarely transcend the paragraph level.

The direct assessment scoring model appeals to our notion of common sense, and is therefore trusted. The obvious disadvantages are the cost and time required to score the essays. In addition, the reliability of the scores is a potential problem, although there is literature available that describes measures to increase reliability.

Scoring models 1 and 2 produce similar results. The majority of examinees pass both subtests using model 1, and pass by the combined score using model 2. A few examinees who failed one of the two subtests passed by the combined score. This occurred when an above-minimum performance on one subtest pulled up their combined score. Conversely a few examinees who passed one subtest failed by the combined score. This occurred when their performance on one subtest was insufficiently strong to compensate for a weak performance on the

other subtest.

Since scoring models 1 and 2 produce similar results, it is legitimate to ask, "Which scoring model should be used?" or "Does it matter which scoring model is used?" The answer to these questions should not depend upon the pass/fail results produced by the model. The answers should depend upon the nature of the construct being measured. If it is believed that there is one construct, "writing ability," then it is appropriate to measure it with multiple measures, and report a composite score. If there is more than one construct being measured, then it is inappropriate to report only a composite score.

The purpose of the assessment is another factor to be considered when choosing the scoring model. If the purpose of the assessment is to assure that students have acquired the writing skills necessary to compose an essay, then assigning a separate score to the writing sample is the most direct scoring model to use. If, however, mastery of a content area is being measured, then a composite score is easier to interpret and more inclusive.

If the composite scoring model is selected, and one therefore assumes that one construct is being measured, then it is reasonable to expect a strong correlation between the two measures. This expectation was checked (see Table 14) by calculating the Pearson product-moment correlation between the essay subtest and the multiple-choice writing subtest. A cursory glance at Table 14 might lead to the conclusion that our correlations are not impressive. However, it is important to remember that when the population is homogeneous the correlations decrease. Our population (college students near the end of their sophomore year) are sufficiently homogeneous to explain this effect. Therefore we concluded that the correlations had sufficient magnitude to justify the composite scoring model if one assumes that one construct is

being measured by both subtests.

The regression scoring model did not work well with our data because the essay ratings are discrete, restricted data points. It is difficult to find a regression line of best fit because there are restricted values on each axis. In our case, a 4-paper (that passed) was inappropriately shifted to a 3.96 paper (that failed).

However, if we consider the transformed essay scale as a continuous scale, then 3.5 or above could be considered a passing score. In this case, pass/fail decisions would be the same for the converted scores as for the original essay scores.

Two conditions should be present in order to apply the regression model. First, an adequate number of score points are needed to find the line of best fit. We concluded that four discrete score points are inadequate. Secondly, the purpose of the assessment should be to spread out examinee scores across a continuum. This will allow for minor corrections for essay topic difficulty which will change the rank-order of examinees, but not dramatically reverse individual pass/fail decisions.

Further Research

Because important educational decisions are made based upon the scores from large-scale writing assessments, the need for more research in this area is self-evident. Two areas of inquiry are proposed:

1) A generalizability (G) study using analysis of variance to identify the variance due to topic and raters could contribute directly to an improved analysis of the field-test results of prompts, and to improvement in rater reliability.

2) A study using item response theory with polychotomous scores could

facilitate the development of a bank of essay topics with varying difficulty. More able students could be challenged by more difficult topics, whereas less able students could develop their skills using easier topics. This kind of research could lead to computer-generated essay topics to provide tailored testing.

In order for this kind of research to achieve optimal results, it is important that the educational researchers and writing faculty work together so that their efforts are harmonious.

APPENDIX

TABLE 1

DESCRIPTIVE STATISTICS FOR ESSAY EXAMINATIONS

Administration Date: June 1983

	Topic One				Topic Two			
With Referee	Mean	Standard Deviation	Alpha	N	Mean	Standard Deviation	Alpha	N
All Students	4.882	1.521	.799	2982	4.686	1.450	.787	7383
Male	4.526	1.520	.797	1311	4.528	1.431	.780	3511
Female	5.162	1.463	.785	1671	4.830	1.453	.789	3872
White	5.148	1.418	.763	2281	4.982	1.348	.744	5213
Black	4.098	1.396	.795	143	3.716	1.280	.753	617
Hispanic	4.258	1.508	.799	271	4.259	1.480	.800	932
Am.Indian/ Alaskan	4.429	2.225	.962	7	4.625	1.586	.795	16
Asian/ Pacific	4.000	1.661	.867	30	3.688	1.613	.855	48
Without Referee	Mean	Standard Deviation	Alpha	N	Mean	Standard Deviation	Alpha	N
All Students	4.870	1.489	.710	2982	4.684	1.427	.718	7383
Male	4.522	1.487	.704	1311	4.525	1.411	.712	3511
Female	5.142	1.433	.691	1671	4.827	1.426	.718	3872
White	5.129	1.381	.649	2281	4.974	1.326	.664	5213
Black	4.084	1.387	.751	143	3.739	1.277	.669	617
Hispanic	4.295	1.496	.716	271	4.256	1.443	.727	932
Am.Indian/ Alaskan	4.429	2.225	.962	7	4.750	1.613	.694	16
Asian/ Pacific	3.900	1.561	.774	30	3.688	1.613	.855	48

Total Tested 10365
N-Per Cent Splits 420 (4.05%)
N-Per Cent Total Agreement 5520 (53.26%)
Mean Rating 2.37

Per Cent Essays Receiving Each Score

Score: 4 - 7.45%
3 - 35.89%
2 - 42.98%
1 - 13.67%

TABLE 2

DESCRIPTIVE STATISTICS FOR ESSAY EXAMINATIONS

Administration Date: October 1983

	Topic One				Topic Two			
With Referee	Mean	Standard Deviation	Alpha	N	Mean	Standard Deviation	Alpha	N
All Students	4.644	1.343	.759	9546	4.882	1.494	.791	4763
Male	4.376	1.318	.752	4522	4.621	1.505	.797	2135
Female	4.885	1.320	.748	5024	5.094	1.452	.777	2628
White	4.870	1.259	.720	7064	5.132	1.390	.752	3678
Black	3.947	1.277	.745	754	3.989	1.479	.814	261
Hispanic	4.314	1.335	.760	673	4.346	1.449	.796	318
Am.Indian/ Alaskan	4.875	1.116	.665	24	4.889	1.711	.777	18
Asian/ Pacific	4.065	1.681	.874	77	4.143	1.621	.764	21
Without Referee	Mean	Standard Deviation	Alpha	N	Mean	Standard Deviation	Alpha	N
All Students	4.648	1.329	.701	9546	4.881	1.467	.714	4763
Male	4.380	1.301	.689	4522	4.619	1.472	.714	2135
Female	4.890	1.307	.691	5024	5.094	1.428	.699	2628
White	4.871	1.243	.652	7064	5.128	1.360	.654	3678
Black	3.950	1.269	.709	754	3.978	1.470	.769	261
Hispanic	4.324	1.320	.705	673	4.365	1.425	.699	18
Am.Indian/ Alaskan	4.875	1.116	.665	24	4.889	1.711	.777	18
Asian/ Pacific	4.091	1.695	.859	77	4.143	1.621	.764	21

Total Tested	14309	Per Cent Essays Receiving Each Score
N-Per Cent Splits	465 (3.25%)	Score: 4 - 6.36%
N-Per Cent Total Agreement	7757 (54.21%)	3 - 35.58%
		2 - 45.89%
Mean Rating	2.36	1 - 12.16%

TABLE 3

DESCRIPTIVE STATISTICS FOR ESSAY EXAMINATIONS

Administration Date: March 1984

	Topic One				Topic Two			
With Referee	Mean	Standard Deviation	Alpha	N	Mean	Standard Deviation	Alpha	N
All Students	5.112	1.420	.805	8024	4.968	1.471	.826	10317
Male	4.777	1.418	.821	3208	4.713	1.448	.830	5249
Female	5.335	1.376	.781	4816	5.232	1.449	.811	5068
White	5.315	1.309	.755	6187	5.199	1.337	.774	8173
Black	4.318	1.550	.875	751	4.015	1.515	.880	621
Hispanic	4.680	1.451	.833	456	4.531	1.589	.872	558
Am.Indian/ Alaskan	4.842	1.385	.805	19	4.969	1.596	.843	32
Asian/ Pacific	5.108	1.480	.866	65	3.876	1.670	.908	89
Without Referee	Mean	Standard Deviation	Alpha	N	Mean	Standard Deviation	Alpha	N
All Students	5.101	1.388	.701	8024	4.973	1.425	.709	10317
Male	4.760	1.379	.702	3208	4.718	1.400	.699	5249
Female	5.328	1.346	.679	4816	5.237	1.402	.699	5068
White	5.301	1.289	.649	6187	5.189	1.316	.656	8173
Black	4.332	1.479	.751	751	4.079	1.399	.714	621
Hispanic	4.669	1.396	.700	456	4.548	1.516	.745	558
Am.Indian/ Alaskan	4.842	1.302	.646	19	4.938	1.625	.733	32
Asian/ Pacific	5.077	1.407	.647	65	3.899	1.596	.821	89

Total Tested	18341	Per Cent Essays Receiving Each Score
N-Per Cent Splits	678 (3.70%)	Score: 4 - 9.56%
N-Per Cent Total Agreement	9684 (52.80%)	3 - 41.46%
		2 - 39.92%
Mean Rating	2.52	1 - 9.06%

TABLE 4

DESCRITPIVE STATISTICS FOR ESSAY EXAMINATIONS

Administration Date: June 1984

	Topic One				Topic Two			
With Referee	Mean	Standard Deviation	Alpha	N	Mean	Standard Deviation	Alpha	N
All Students	4.854	1.676	.883	2159	5.002	1.463	.828	8108
Male	4.670	1.626	.880	1245	4.749	1.440	.830	3456
Female	5.104	1.712	.882	914	5.191	1.452	.819	4652
White	5.374	1.446	.813	1422	5.258	1.339	.776	6113
Black	3.859	1.565	.897	234	4.044	1.442	.881	544
Hispanic	4.319	1.569		157	4.581	1.457	.862	676
Am.Indian/ Alaskan	4.500	2.082		4	4.957	1.397	.852	23
Asian/ Pacific	3.708	2.177	.981	24	4.269	1.586	.854	52
Other	3.613	1.620	.931	318	3.977	1.638	.907	700
Without Referee	Mean	Standard Deviation	Alpha	N	Mean	Standard Deviation	Alpha	N
All Students	4.850	1.632	.797	2159	4.985	1.423	.726	8108
Male	4.659	1.583	.788	1245	4.729	1.402	.716	3456
Female	5.109	1.663	.799	914	5.175	1.409	.722	4652
White	5.350	1.441	.733	1422	5.228	1.321	.674	6113
Black	3.855	1.455	.753	234	4.063	1.358	.725	544
Hispanic	4.389	1.489	.778	157	4.573	1.422	.769	676
Am.Indian/ Alaskan	4.250	1.708	.490	4	4.913	1.443	.834	23
Asian/ Pacific	3.750	2.090	.913	24	4.250	1.413	.684	52
Other	3.664	1.539	.780	318	4.033	1.528	.764	700

Total Tested	10268
N-Per Cent Splits	328 (3.19%)
N-Per Cent Total Agreement	5589 (54.43%)
Mean Rating	2.49

Per Cent Essays Receiving Each Score

Score	
4	9.75%
3	39.88%
2	39.51%
1	10.86%

TABLE 5

DESCRIPTIVE STATISTICS FOR ESSAY EXAMINATIONS

Administration Date: September 1984

	Topic One				Topic Two			
With Referee	Mean	Standard Deviation	Alpha	N	Mean	Standard Deviation	Alpha	N
All Students	4.961	1.413	.814	8745	4.826	1.410	.819	4233
Without Referee								
All Students	4.942	1.375	.700	8745	4.798	1.376	.705	4233

Total Tested	13868	Per Cent Essays Receiving Each Score	
N-Per Cent Splits	492 (3.55%)	Score	4 - 7.74%
N-Per Cent Total Agreement	7581 (54.67%)		3 - 37.86%
			2 - 43.88%
Mean Rating	2.43		1 - 10.52%

TABLE 6

DESCRIPTIVE STATISTICS FOR ESSAY EXAMINATIONS*

Administration Date: March 1985

	Topic One				Topic Two			
With Referee	Mean	Standard Deviation	Alpha	N	Mean	Standard Deviation	Alpha	N
All Students	5.129	1.296	.760	8399	5.152	1.535	.835	5259
Without Referee								
All Students	5.102	1.268	.649	8399	5.132	1.495	.742	5259

Total Tested	16067	Per Cent Essay Receiving Each Score	
N-Per Cent Splits	515 (3.21%)	Score	4 - 9.82%
N-Per Cent Total Agreement	8611 (53.59%)		3 - 40.70%
			2 - 40.23%
Mean Rating	2.51		1 - 9.25%

No breakdown by sex or race/ethnic group was available for these administrations.

TABLE 7

Difference in Means for Essay Topics
(Topic One - Topic Two)

With Referee	6/83	10/83	3/84	6/84	9/84*	3/85*
All Students	.196	-.238	.144	-.148	.135	-.023
Male	-.002	-.245	.064	-.079		
Female	.332	-.209	.103	-.087		
White	.166	-.262	.116	.116		
Black	.382	-.042	.303	-.185		
Hispanic	-.001	-.032	.149	-.262		
Am.Indian	-.196	-.014	-.127	-.457		
Asian	.312	-.078	1.232	-.561		
Other				-.364		

TABLE 8

Difference in Alpha for Essay Topics
(Topic One - Topic Two)

With Referee	6/83	10/83	3/84	6/84	9/84*	3/85*
All Students	.012	-.032	-.021	.055	-.005	-.075
Male	.017	-.045	-.009	.050		
Female	-.004	-.029	-.030	.063		
White	.019	-.032	-.019	.037		
Black	.042	-.069	-.005	.016		
Hispanic	-.001	-.036	-.039			
Am.Indian	.167	-.112	-.038			
Asian	.012	.110	-.042	.127		
Other				.024		

TABLE 9

Percent Choosing Each Topic

	6/83	10/83	3/84	6/84	9/84	3/85
Topic 1	.29	.67	.44	.21	.61	.67
Topic 2	.71	.33	.56	.79	.39	.33

*No breakdown by sex or race/ethnic group was available for these administrations.

TABLE 10

Descriptive Statistics for Essay and Writing Subtests

	Essay Mean	Essay Std Dev.	Number Examinees	Writing Mean	Writing Std Dev.	Number Examinees
March '83	4.88	1.45	19,044	307.24	30.91	19,061
June '83	4.74	1.47	10,365	303.51	30.66	10,376
October '83	4.72	1.40	14,309	307.23	32.40	14,323
March '84	5.03	1.45	18,343	314.05	29.67	18,339
June '84	4.97	1.51	10,268	309.68	30.08	10,191
September '84	4.86	1.44	13,870	319.23	33.14	13,605

TABLE 11

Descriptive Statistics for Combined Score

	Mean	Std Dev.	Number
March '83	603.68	64.59	19,037
June '83	595.87	64.91	10,356
October '83	599.01	64.66	14,299
March '84	614.98	63.82	18,334
June '84	609.11	65.71	10,190
September '84	616.59	65.83	13,462

TABLE 12

Examinees Passing Essay and Writing Subtests

Total N		Number (%) Passing Essay	Writing	Both	Combined
19,070	March '83	15,951(83.6%)	17,948(94.1%)	15,564(81.6%)	16,297(85.5%)
10,389	June '83	8,398(80.8%)	9,607(92.5%)	8,104(78.0%)	8,534(82.1%)
14,339	Oct '83	11,809(82.4%)	13,190(92.0%)	11,398(79.5%)	12,032(83.9%)
18,352	March '84	16,665(90.8%)	17,621(96.0%)	16,416(89.5%)	16,583(90.4%)
10,270	June '84	9,153(89.1%)	9,725(94.7%)	8,908(86.7%)	8,974(87.4%)
14,431	Sept '84	12,410(86.0%)	13,126(91.0%)	11,986(83.1%)	12,105(83.9%)

TABLE 13

Examinees Passing by Combined Scores*
but Failing Essay or Writing

	Pass Both	Passed Combined Failed Other	Failed Writing	Failed Essay	Total Tested
March '83	15,564(81.6%)	733(3.8%)	138(.7%)	595(3.0%)	19,070
June '83	8,104(78.0%)	430(4.1%)	119(1.2%)	311(3.0%)	10,389
Oct '83	11,398(79.5%)	634(4.4%)	129(.9%)	505(3.5%)	14,339
March '84	16,416(89.5%)	168(.9%)	84(.5%)	83(.5%)	18,352
June '84	8,908(86.7%)	66(.6%)	50(.5%)	16(.2%)	10,220
Sept '84	11,986(83.1%)	119(.8%)	73(.5%)	46(.3%)	14,431

*No one passed combined and failed both essay and writing.

TABLE 14

Correlations of Essay and Writing Scores

	PPM Corr.	N
March '83	.49	19,037
June '83	.49	10,356
Oct '83	.51	14,299
March '84	.50	18,334
June '84	.50	10,190
Sept '84	.50	13,462

TABLE 15

Descriptive Statistics for Essay and Writing Subtests by Topic

		Essay Mean	Essay Std Dev.	Number Examinees	Writing Mean	Writing Std Dev.	Number Examinees
Mar. 83	Topic 1	4.79	1.39	15,583	307.37	30.76	15,577
	Topic 2	5.29	1.65	3,461	306.86	31.18	3,460
June 83	Topic 1	4.88	1.52	2,982	306.53	30.51	2,980
	Topic 2	4.69	1.45	7,383	302.34	30.59	7,376
Oct. 83	Topic 1	4.64	1.34	9,546	305.77	31.51	9,537
	Topic 2	4.88	1.49	4,763	310.40	33.65	4,762
Mar. 84	Topic 1	5.11	1.42	8,024	315.59	29.64	8,021
	Topic 2	4.97	1.47	10,317	312.86	29.63	10,311
June 84	Topic 1	4.85	1.68	2,159	306.03	32.08	2,138
	Topic 2	5.00	1.46	8,109	310.65	29.45	8,052
Sept 84	Topic 1	4.90	1.45	9,314	319.28	32.89	9,031
	Topic 2	4.77	1.43	4,554	320.19	33.30	4,429

TABLE 16

Percent of Examinees Passing Based on Writing and Essay Subtest, and by a Combined Score, by Topic

		Pass Essay	Pass Writing	Pass Combined
Mar. 83	Topic 1	83.5	94.2	85.4
	Topic 2	85.0	94.0	86.4
June 83	Topic 1	82.3	93.6	84.0
	Topic 2	80.5	92.1	81.6
Oct. 83	Topic 1	82.3	91.8	83.7
	Topic 2	83.0	92.6	84.9
Mar. 84	Topic 1	92.1	96.6	91.7
	Topic 2	89.9	95.6	89.4
June 84	Topic 1	84.3	91.4	81.9
	Topic 2	90.4	95.6	88.9
Sept 84	Topic 1	89.9	93.8	87.7
	Topic 2	88.7	94.0	86.5

TABLE 17

Adjusted Means and Standard Deviations

	ESSAY				WRITING		
	Mean	Std Dev.	Adjusted Mean	Adjusted Std Dev.	Mean	Std Dev.	N
March 1983							
Topic One	4.789	1.390	-	-	307.371	30.755	15,577
Topic Two	5.292	1.645	4.777	1.395	306.859	31.184	3,460
June 1983							
Topic One	4.882	1.521	4.769	1.388	306.531	30.505	2,982
Topic Two	4.686	1.450	4.676	1.388	302.344	30.587	7,383
Oct. 1983							
Topic One	4.644	1.343	4.753	1.399	305.767	31.505	9,546
Topic Two	4.882	1.494	4.856	1.424	310.395	33.653	4,763

TABLE 18

Rescaled Scores for Examinees

Essay Score:	2	3	4	5	6	7	8
March 1983							
Topic One	2	3	4	5	6	7	8
Topic Two	1.985	2.833	3.681	4.529	5.377	6.225	7.073
June 1983							
Topic One	2.141	3.053	3.965	4.877	5.790	6.702	7.614
Topic Two	2.104	3.061	4.019	4.977	5.934	6.892	7.849
Oct. 1983							
Topic One	1.999	3.040	4.082	5.123	6.165	7.206	8.248
Topic Two	2.110	3.063	4.016	4.969	5.922	6.874	7.827

TABLE 19

Percent of Examinees Passing with Rescaled Scores at a Cutting Score of 4

	PERCENT PASSING		
March 1983	Without Equating	With Equating	N
Topic One	.84	.84	15,577
Topic Two	.85	.69	3,460
June 1983			
Topic One	.82	.60	2,982
Topic Two	.80	.80	7,383
Oct. 1983			
Topic One	.82	.82	9,546
Topic Two	.83	.83	4,763

REFERENCES

Baron, J. B. Writing assessment in Connecticut: a holistic eye toward identification and an analytic eye toward instruction. Educational Measurement: Issues and Practice, 1984, 3 (1), 27-28; 38.

Breland, H. M. The direct assessment of writing skill: a measurement review (College Board Report No. 83-6). Princeton, N. J.: College Entrance Examination Board, 1983.

Chapman, C. W., Fyans, L. J., Jr., & Kerins, C. T. Writing Assessment in Illinois. Educational Measurement: Issues and Practice, 1984, 3 (1), 24-26.

College Entrance Examination Board. College-level examination program: what your scores mean. New York: Author, 1981.

College Entrance Examination Board. Guide to CLEP examinations. New York: Author, 1977.

College Entrance Examination Board. Guide to examinations in literature. New York: Author, 1977.

Crocker, L., & Algina, J. Introduction to classical and modern test theory. New York: Holt, Renehart, & Winston, 1986.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: a liberalization of reliability theory. The British Journal of Statistical Psychology, 1963, 16 (2), 137-163.

De Gruijter, D. N. M. The essay examination. In L. J. Th. van der Kamp, W. F. Langerak & D. N. M. de Gruijter (Eds.), Psychometrics for educational debates. London: Wiley, 1980.

Diederich, P. B. Measuring growth in English. Urbana, Ill.: National Council of Teachers of English, 1974.

Freedman, S. W. & Robinson, W. S. Testing proficiency in writing at San Francisco State University. College Composition and Communication, 1982, 37, 393-398.

Godshalk, F., Swineford, F., & Coffman, W. The measurement of writing ability. New York: College Entrance Examination Board, 1966.

Hermann, F. & Williams, P. L. Writing assessment in Maryland. Educational Measurement: Issues and Practice, 1984, 3 (1), 23-24.

Hills, J. R. Consistent college grading standards through equating. Educational and Psychological Measurement, 1972, 32, 137-146.

Hoetker, J. Essay examination topics and students' writing. College Composition and Communication, 1982, 37, 377-392.

Kane, M. T., Gillmore, G. M. & Crooks, T. J. Student evaluations of teaching: the generalizability of class means. Journal of Educational Measurement, 1976, 13, 171-183.

Meredith, V. H. Writing assessment in South Carolina: past and present. Educational Measurement: Issues and Practice, 1984, 3 (1), 19-21.

Meredith, V. H. & Williams, P. L. Issues in direct writing assessment: problem identification and control. Educational Measurement: Issues and Practice, 1984, 3 (1), 11-15; 35.

Meyer, G. The choice of questions on essay examinations. Journal of Educational Psychology, 1939, 30, 161-171.

Mullis, I. Scoring direct writing assessments: what are the alternatives? Educational Measurement: Issues and Practice, 1984, 3 (1), 16-18.

Quellmalz, E. S. Toward successful large-scale writing assessment: where are we now? Where do we go from here? Educational Measurement: Issues and Practice, 1984, 3 (1), 29-32; 35.

Rosenbaum, P. R. Model-based direct adjustment. Princeton, N. J.: Educational Testing Service, 1985.

Sachse, P. P. Writing assessment in Texas: practices and problems. Educational Measurement: Issues and Practice, 1984, 3 (1), 21-23.

Williams, P. L. The application of direct writing assessments in five states. Educational Measurement: Issues and Practice, 1984, 3 (1), 19.

Wood, R. & Wilson, D. T. Determining a rank order when not all individuals are assesed on the same basis. In L. J. Th. van der Kamp, W. F. Langerak & D. N. M. de Gruijter (Eds.), Psychometrics for educational debates. London: Wiley, 1980.

REFERENCE NOTES

1. Boodoo, G. M. Inter-rater reliability in generalizability theory. Paper presented at the annual meeting of the National Council on Measurement in Education held in New Orleans, April, 1984.

2. Shavelson, R. J. Application of generalizability theory to measurements from hierarchical populations. Paper presented at the annual meeting of the American Educational Research Association held in New Orleans, April, 1984.

3. Williams, P. L. Practical applications of item response theory in large scale testing programs: the domain score estimation of equating direct writing assessments. Paper presented at the 12th Annual Large Scale Assessment Conference held in Boulder, Colorado, June, 1983.